Neuropsychologia 112 (2018) 77–85

Contents lists available at ScienceDirect

Neuropsychologia

journal homepage: www.elsevier.com/locate/neuropsychologia

The two processes underlying the testing effect– Evidence from Event-Related Potentials (ERPs)

Xiaonan L. Liu[a,b,c,*], Deborah H. Tan[b,d], Lynne M. Reder[b,c]

[a] Institute of Psychology, School of Public Policy, Xiamen University, Xiamen, China 361005
[b] Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[c] Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[d] Department of Psychology, University of Minnesota, Minneapolis, MN 55455, USA

ARTICLE INFO

Keywords:
Testing effect
Retrieval attempt
Re-encoding
Event-Related Potentials (ERP)
Cued recall

ABSTRACT

Theoretical explanations of the testing effect (why people learn better from a test than a re-study) have largely focused on either the benefit of attempting to retrieve the answer or on the benefit of re-encoding the queried information after a successful retrieval. While a less parsimonious account, prior neuroimaging evidence has led us to postulate that both of these processes contribute to the benefit of testing over re-study. To provide further empirical support for our position, we recorded ERPs while subjects attempted to recall the second word of a pair when cued with the first. These ERPs were analyzed based on the current response accuracy and as a function of accuracy on the subsequent test, yielding three groups: the first and second tests were correct, the first was correct and the second was not, both were incorrect. Mean amplitude waveforms during the first test showed different patterns depending on the outcome patterns: Between 400 and 700 ms the amplitudes were most positive when both tests were correct and least positive when both were incorrect; mean amplitudes between 700 and 1000 ms only differed as a function of *subsequent memory*. They were more positive when the second test was correct. Importantly, the later component only predicted subsequent memory when the answers were not overlearned, i.e. only correctly recalled once previously. We interpret the 400–700 ms time window as a component reflecting a retrieval attempt process, which differs as a function of both current and subsequent accuracy, and the later time window as a component reflecting a re-encoding process, which only involves learning from tests, both of which are involved in the testing effect.

1. Introduction

There is ample evidence that learning procedures that involve testing are more effective than procedures that only involve re-study. This result has been referred to as the *Testing Effect* or the effect of Retrieval Practice (Karpicke and Roediger, 2008; Roediger and Butler, 2011; Roediger and Karpicke, 2006a, b). A typical memory experiment that explores the Testing Effect involves an initial study phase in which all items are presented in the same manner but then are practiced either through additional study trials (re-study condition) or through testing (test condition). The typical finding of this paradigm is that, on a final memory test, items practiced in the test condition are remembered better than those practiced in the re-study condition (Roediger and Karpicke, 2006b; Toppino and Cohen, 2009).

While the testing effect has been rigorously studied (Hogan and Kintsch, 1971; Pyc and Rawson, 2009; Wheeler and Roediger, 1992), there have been surprisingly few mechanistic accounts for the phenomenon. Some of the contemporary explanations of the test advantage focus on the retrieval processes involved, while others focus on the post-retrieval re-encoding process underlying the testing effect. Importantly, these theoretical explanations have tended to focus on only one of the processes. For example, the Elaborative Retrieval Account focuses on the retrieval process whereby a search is initiated to find the answer to the question. This theory states that the retrieval of information from memory results in memory elaboration and/or in forming new associations to the correct answers, which makes the information more likely to be successfully retrieved again in the future (Anderson and Reder, 1979; Carpenter, 2009; Carpenter and Delosh, 2006). Another example of a theory focusing on this process is the Episodic Context Account proposed by Karpicke et al. (2014). This account states that retrieval serves to add unique contextual information to the memory trace, making subsequent retrieval easier. On the other hand, the Reconsolidation Account (Finn and Roediger, 2011) emphasizes a re-encoding process postulated to occur after retrieval (i.e.,

* Correspondence to: Institute of Psychology, School of Public Policy, Xiamen University, Fujian, China, 361005.
E-mail address: xliu@xmu.edu.cn (X.L. Liu).

https://doi.org/10.1016/j.neuropsychologia.2018.02.022
Received 3 August 2017; Received in revised form 19 February 2018; Accepted 19 February 2018
Available online 21 February 2018
0028-3932/

when the correct answers are in working memory.) This account claims that after the first successful retrieval of the studied information, the retrieved information enters an unstable state (Dudai, 2004), thereby enabling the memory trace to be strengthened by the post-retrieval re-encoding of the correctly retrieved information.

Each type of theory provides a plausible explanation of the testing effect using either the retrieval or the re-encoding process. However, one might wonder whether both processes are involved.

Recent neuroimaging evidence (Liu et al., 2014; Liu and Reder, 2016) provides support for this point of view. Unlike previous studies in which subsequent memory analyses had been used to analyze *encoding trials* back-sorted on whether the material was later successfully remembered (Wagner et al., 1998), Liu et al. (2014) and Liu and Reder (2016) employed a subsequent memory analysis on *test trials* (i.e., retrieval practice), and uncovered two sets of neural processes involved in the testing effect: a retrieval process that involves attempting to retrieve the answer, and a re-encoding process that involves re-encoding the answer that had just been retrieved. Those studies found that brain regions in the left hemisphere, including the prefrontal cortex (PFC), posterior parietal cortex (PPC) and hippocampus (HPC), previously shown to predict subsequent memory performance based on activity during encoding (Kim, 2011), were also predictive of subsequent performance based on activity during test trials. On the other hand, other regions in the right hemisphere (the right PFC and right PPC) were found to be predictive of subsequent memory performance only when subsequent memory analyses were done on test trials (Liu et al., 2014).

These two patterns of results led Liu et al. (2014) and Liu and Reder (2016) to conjecture that there are two processes underlying the benefit of testing: the retrieval process, which involves the regions in the right hemisphere, and the re-encoding process, which involves the regions in the left hemisphere. Other fMRI studies that have employed subsequent memory analyses on test and re-study trials have similarly found that learning through testing involves additional brain regions compared with those active through re-study opportunities, suggesting that additional processes are involved during testing compared to re-study. (Den Broek et al., 2013; Wing et al., 2013).

Our account of the testing effect advantage involves postulating two separate processes during a testing episode that necessarily occur sequentially: first the retrieval of the answer, followed by the re-encoding of that answer. One limitation of Liu et al. (2014) and Liu and Reder (2016) was that their evidence was based solely on fMRI data, which has poor temporal sensitivity. In order to provide converging support that will also provide strong temporal evidence, we chose to employ ERP (Event Related Potential) methodology to illustrate that these two processes are dissociable in time. EEG provides excellent temporal resolution and should allow us to determine whether there are two temporally distinct processes and the temporal order of the two proposed processes. If our theory is correct, we should be able to identify multiple ERP components corresponding to the two processes proposed that are predictive of subsequent test performance.

Few studies have used ERPs to investigate mechanisms underlying the testing effect. Most of them focused on recording ERPs at the final test (e.g., Spitzer et al., 2009; Rosburg et al., 2015). One notable exception is the study by Bridge and Paller (2012) who also directly examined the ERPs associated with intervening tests. Bridge and Paller (2012) asked subjects to learn to associate objects with locations and then asked subjects to recall the location when presented with the object. There were multiple rounds of testing in which the investigators examined amplitude differences at two time windows, 400–700 ms and 700–1000 ms as a function of accuracy of response and ability to predict whether the answer would be correct on the subsequent test. They found that during the tests, the amplitude of the ERPs at the earlier time window (400–700 ms) was more positive the closer the recalled location was to the correct location. The amplitude of ERPs during the second time window (700–1000 ms), however, was not correlated with location accuracy on the current test. This amplitude was instead (and surprisingly) correlated with performance on the *subsequent test*: amplitude of waveforms during the second time window were positively correlated with the proximity of the location recalled on the current test to the location recalled on the subsequent test. That is, while the amplitude of the waveform in the second time window did not predict accuracy of the current test response, the more positive the amplitude, the closer the location of the subsequent answer was to the current location answer. In other words, the late component reflects the encoding of what is retrieved on the current test, similar to what we conjectured in our fMRI paper (Liu and Reder, 2016).

In the current study, we test the hypothesis that there are two memorial processes associated with the testing effect. We do this by examining two ERP time windows: the first, between 400 and 700 ms, we associate with attempting to retrieve the answer; the second, between 700 and 1000 ms, we assume is associated with re-encoding the answer. If our hypothesis is correct, then we should observe two different ERP effects: the amplitudes of both time windows during recall should predict subsequent memory performance, but only the amplitudes between 400 and 700 ms should reflect current accuracy.

It is worth noting that the choice of time windows is also consistent with prior research (e.g., Allan and Rugg, 1997; Bai et al., 2015; Wilding and Rugg, 1996; Wilding and Ranganath, 2012; Curran, 2000; Johansson and Mecklinger, 2003). Allan and Rugg (1997) examined the retrieval success effect of a cued-recall task and found that successful retrieval elicited a more positive shift compared with the control task. In addition, this shift was evident around 400 ms after the stimulus appeared. Other studies (e.g., Curran, 2000; Johansson and Mecklinger, 2003; Bai et al., 2015) examined a later component that was associated with post-retrieval assessment and monitoring, and indicated that this component was usually not observed before 700 ms after onset of the stimuli.

1.1. Why we chose no feedback in this experiment

The testing effect is an important pedagogical result that is usually found in studies that provide feedback after each testing attempt (Roediger and Karpicke, 2006a, 2006b; Kang et al., 2007) or until the tests are accurate (Karpicke and Roediger, 2008). In our first study (Liu et al., 2014), we too used feedback after each test; however, in order to more carefully test our theory of the advantage derived from testing over re-study, it was important to also run experiments in which no feedback was provided. The first of the two postulated processes, retrieval, can be observed regardless of whether feedback is given, but the second postulated process of re-encoding the answer could conceivably be compromised by the presentation of feedback. Thus, in order to isolate re-encoding that is based on retrieval as compared with re-encoding from feedback, we did not provide feedback after testing.

2. Materials and methods

2.1. Subjects

Thirty-one subjects (19 females, mean age 19 ± 2.01) with normal or corrected-to-normal vision participated in two sessions with one day between them. All subjects were students studying at Carnegie Mellon University. Subjects were paid $20 after completion of both sessions. All subjects were treated in accordance with the CMU IRB guidelines.

2.2. Design, materials and procedure

Fig. 1 illustrates the various conditions included in this within-subject design. All conditions involved an initial encoding (study) phase (Phase 1). The top row illustrates one of the conditions: Study followed by a Test, a second Test and a third Test, denoted as STTT. Word pairs in this condition were tested twice on the first day after the initial encoding period. The final test occurred on the second day, and was

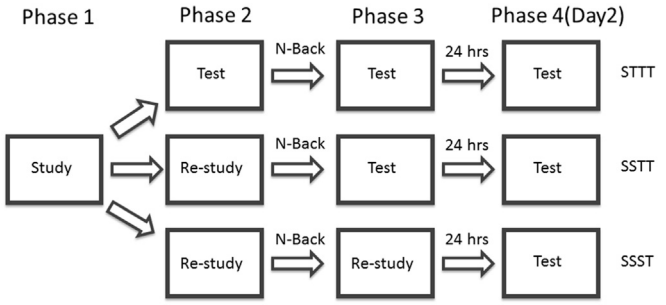

Fig. 1. This is a within-subject design. The three rows illustrate the three treatment conditions for word-pairs on a single list. Phases 1, 2, and 3 occurred on Day 1. Phase 4 occurred on Day 2. ERPs were only collected on Day 1.

called Phase 4. As depicted in Fig. 1, all pairs from all conditions were tested on Day 2 (Phase 4). The middle row of Fig. 1 illustrates that the pairs in the SSTT condition were re-studied once after the initial encoding (two study opportunities in all), and this was followed by a test on Day 1 as well as the final test on Day 2. The bottom row of Fig. 1 illustrates the pairs in the SSST condition, which were re-studied two more times following the initial encoding on Day 1. These pairs were only tested on the final test on Day 2.

For each subject, we randomly selected 420 words from our pool of 480 words to form semantically-unrelated word pairs that were assigned to the three different study/test treatment conditions. The words were selected from the MRC psycholinguistic database (Coltheart, 1981) with the following constraints: 4–7 letters in length and ratings between 500 and 700 for printed familiarity, concreteness, and ease of imagery.

Word pairs in the different conditions were randomly intermixed and assigned to 10 lists for each subject. The only constraints on the random selection and assignment were that no word was used more than once in a given subject's material set and that each of the ten lists consisted of 21 paired associates, 11 for the STTT condition, 6 for the SSTT condition and 4 for the SSST condition.[1] All random assignments were done separately for each subject.

For Phase 1, each word-pair was presented for 3 s before the next pair was shown. The words in a pair were presented side-by-side in the center of the screen. Each study trial began with a fixation cross for a jittered period of 800–1200 ms. After initial study of all pairs for a given list in Phase 1, subjects were given an opportunity to learn each pair once more in Phase 2, with some pairs shown for re-study and some pairs tested for the first time. The order of word pairs in Phase 2 was also randomized. There was no feedback after a recall attempt, as explained in the introduction. After going through Phase 1 and Phase 2 for a specific list, the next list was presented for initial study (Phase 1) followed by its Phase 2 training. Phase 1 and Phase 2 together lasted approximately 30 min.

After all 10 lists had received Phase 1 and Phase 2 training, subjects were given a distractor task (i.e., the N-Back task) for ten minutes prior to starting Phase 3. All 210 word-pairs kept their original treatment condition assignment, but the pairs were randomly re-assigned to different lists with the same constraints as before (balance of the number of pairs from each condition per list). This additional randomization was intended to reduce any noise due to idiosyncratic effects from intralist interactions between word pairs. Phase 3 lasted approximately 20 min.

For Phase 2 and Phase 3, re-study trials were the same as in Phase 1. Test trials also began with a fixation cross for 800–1200 ms, followed by the cue word (the left-hand-side word) in the center of the screen with a question mark prompt to indicate that the subject should try to recall the corresponding right-hand-side word. One second after the onset of the cue and the question mark, the question mark was replaced with an underscore mark indicating that the subject should type out the target. Cued-recall trials were self-paced with a time out after 8 s.

The final assessment (Phase 4) occurred on Day 2. Test trials in Phase 4 were the same as test trials in Phase 2 and 3. Phase 4 lasted approximately 20 min.

2.3. ERP recording

Subjects sat in an electrically-shielded booth. Stimuli were presented on a standard CRT monitor placed behind radio frequency-shielded glass. The CRT monitor was placed approximately 70 cm away from subjects. ERP recordings were made using 32 Ag–AgCl sintered electrodes (10–20 system) and a bio-amplification system (Neuroscan Inc., Sterling, VA). Impedances were adjusted to be less than 5 kΩ. Data were sampled at a rate of 1 kHz with a band pass filter of .1–200 Hz. The left mastoid served as the reference electrode, and scalp recordings re-referenced offline to the average of the right and left mastoids.

2.4. ERP analyses

The EEG recording was decomposed into independent components using the EEGLAB FastICA algorithm (Delorme and Makeig, 2004). Components associated with eye blinks were visually identified and projected out of the EEG recording. ERPLAB (Lopez-Calderon and Luck, 2014) was used for further analyses. The continuous data were segmented from –200–1000 ms relative to trial onset and corrected over the pre-stimulus interval. Trials contaminated with voltages above 100 μV or below − 100 μV were excluded from the analysis. The segmented data were then averaged across trials within each subject for each condition. Based on prior studies (Bridge and Paller, 2012; Bai et al., 2015; Griffin et al., 2013), we focused on two clusters of electrodes – a frontal cluster (F3, Fz, and F4) and a parietal cluster (P3, Pz, and P4)[2] – and two time windows, 400–700 ms and 700–1000 ms. The dependent measures in the ERP analyses were the mean amplitudes of the ERP components in the given time ranges and electrode clusters. Amplitudes were compared using 2 (electrode cluster) X 3 (retrieval outcome patterns) repeated measures analyses of variance (ANOVA). All post-hoc tests were evaluated with a Bonferroni correction to protect against alpha slippage. For plotting, data were first smoothed using a 30 Hz low pass filter.

3. Results and discussion

3.1. Behavioral results

We first analyzed recall accuracy for Phases 3 and 4 as a function of the type of preceding learning experience (the third column of Table 1 and Fig. 2). There were no significant effects of learning experience on accuracy for either phase ($p > .05$). It is likely that this pattern was the result of relatively poor learning during initial study (Phase 1), coupled with a lack of feedback when tested (Kornell et al., 2011; Rowland, 2014). Rowland and Delosh (2015) found that whether the classic testing effect occurs when there is no feedback is moderated by the difficulty of initial retrieval. That is, when the items are more difficult to learn after an initial study, subjects are less likely to get many right on the initial test. As a consequence, there are fewer tacit re-exposures due to fewer successful retrievals. When this happens, performance is better in the re-study condition (e.g., Wheeler et al., 2003). Following

[1] The reason for different numbers of trials in each condition was to make sure there was a sufficient number of subsequently correct and subsequently incorrect trials. This concern is based on the analysis of the second test in the STTT condition for which only trials that were correct on the first test were included.

[2] There was no significant interaction between electrodes within a cluster and retrieval outcome patterns.

Table 1
Mean proportion correct for all test trials regardless of prior accuracy, test trials following correct test(s), and mean number of correct and incorrect items for each phase.[a].

Phase	Condition	All Trials	Trials following correct test (s)	Number of Correct trials	Number of Incorrect trials
Phase 2	sTt-t	.68 (.04)		75	35
Phase 3	stT-t	.52 (.04)	.76 (.02)	57	18
	ssT-t	.59 (.04)		35	25
Phase 4	stt-T	.28 (.03)	.54 (.02)	31	26
	sst-T	.31 (.03)	.52 (.03)	18	17
	sss-T	.29 (.03)			

[a] Standard errors are shown in parentheses. Note that the T that is capitalized refers to the current test phase being analyzed in that row.

compared with when prior learning involved one re-study and one test, $p > .1$. These results are consistent with Rowland and Delosh (2015) in that, when the analyses are conditionalized on initial retrieval success, there is a clear testing effect. This is true for both delay periods: short (10 min, the time between Phase 2 and 3) and long intervals (24 h, the time between Phase 3 and 4).

3.2. ERP results

Our main focus concerns the examination of ERPs during testing. Specifically, we are interested in whether ERPs during a correctly answered test trial will predict whether the next recall of the same item will also be correct (as opposed to the next recall becoming an error). Therefore, different from the behavioral analyses where we used conditionalized analyses, in the ERP results, we analyzed all test trials as a

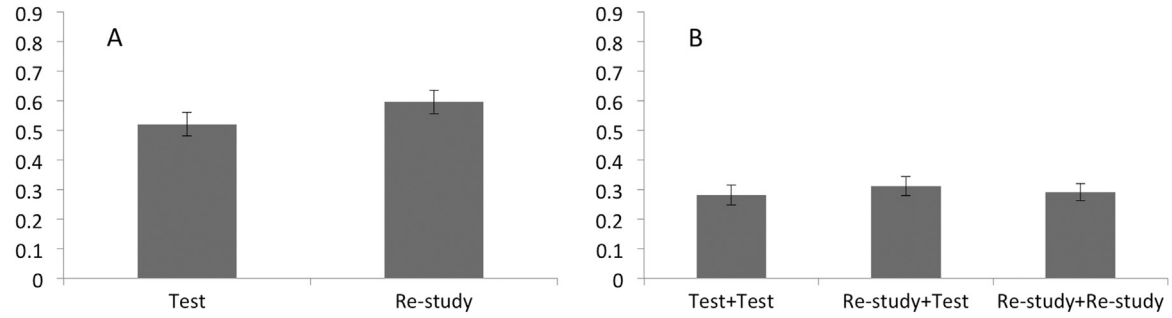

Fig. 2. Mean proportion correctly recalled in Phase 3 (Panel A) and Phase 4 (Panel B), as a function of prior learning experience.

Rowland and Delosh (2015), we also performed analyses conditionalized on initial retrieval success of trials. In the General Discussion and Conclusion, we will discuss in more detail the conditions that produce better performance in testing versus re-study.

The fourth column of Table 1 and Fig. 3A present cued-recall accuracy for Phase 3 for all pairs that were correctly recalled in Phase 2. This column and Fig. 3B also present accuracy for Phase 4 for items correctly recalled in both Phases 2 and 3 and for items re-studied in Phase 2 and correctly recalled in Phase 3. Recall accuracy for Phase 3 was significantly better for trials that followed a correctly recalled test than for trials that followed a re-study trial, $t(30) = 7.34$, $p < .001$, d = 1.31. Recall accuracy for Phase 4 (final assessment on Day 2) showed a significant main effect of type of learning condition, $F(2,60) = 71.13$, $p < .001$, $\eta_p^2 = .70$. A post-hoc test indicated that performance was better for items that had been recalled correctly at least once compared with items that had only been re-studied (sttT vs. sssT: $t(30) = 9.29$, $p < .001$, d = 1.69; sstT vs. sssT: $t(30) = 8.89$, $p < .001$, d = 1.62; the capital letter's position in the string denotes which phase is involved in the comparison). There was no significant difference in recall accuracy in Phase 4 when prior learning opportunities involved two tests

function of current and subsequent recall performance but we did not compare ERPs of test trials with that of re-study trials. Given that the number of observations in a subsequent memory contrast differs as a function of each subject's accuracy, for a specific contrast, we only included subjects that had a minimum of 15 trials per condition (Griffin et al., 2013). The mean numbers of correct and incorrect trials in each condition are presented in Table 1. The total accuracy for that learning condition and accuracy that is based on having been exposed to the correct information in the preceding phase are presented in columns 3 and 4, respectively.

3.2.1. Subsequent memory effects based on ERPs during the first test

In order to determine whether the ERP patterns during the first test would predict current accuracy (on the first test) and future accuracy (on the second test), we examined the ERP patterns for trials in the sTtt condition during Phase 2 and trials in the ssTt condition during Phase 3 (see Fig. 1, top two rows). We examined the two time windows of the retrieval and post-retrieval re-encoding processes (400–700 ms and 700–1000 ms, respectively) as a function of current and future accuracy. Fig. 4 shows topographic maps for the two time windows for each

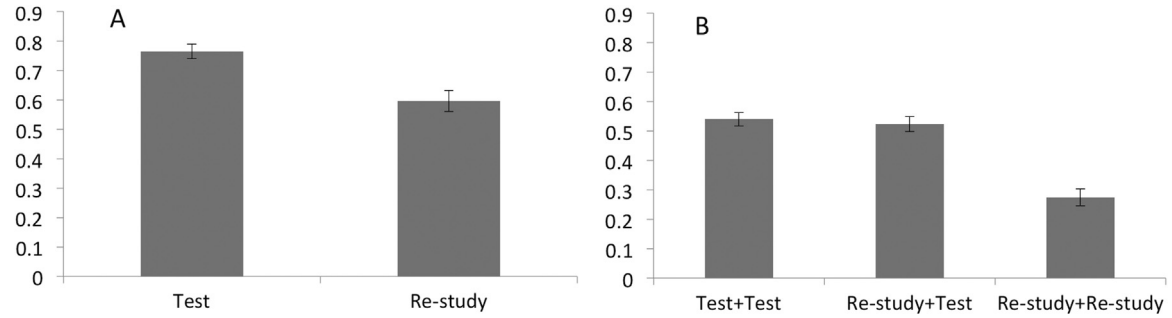

Fig. 3. Mean proportion correctly recalled in Phase 3 (Panel A) and Phase 4 (Panel B), as a function of prior learning experience. These plots include only items that were successfully recalled on the tests that preceded that Phase (3 in Panel A and 4 in Panel B), accuracy for those pairs that had previously been re-studied without any testing, and accuracy for those pairs (in Phase 4) that had been re-studied once and successfully recalled once.

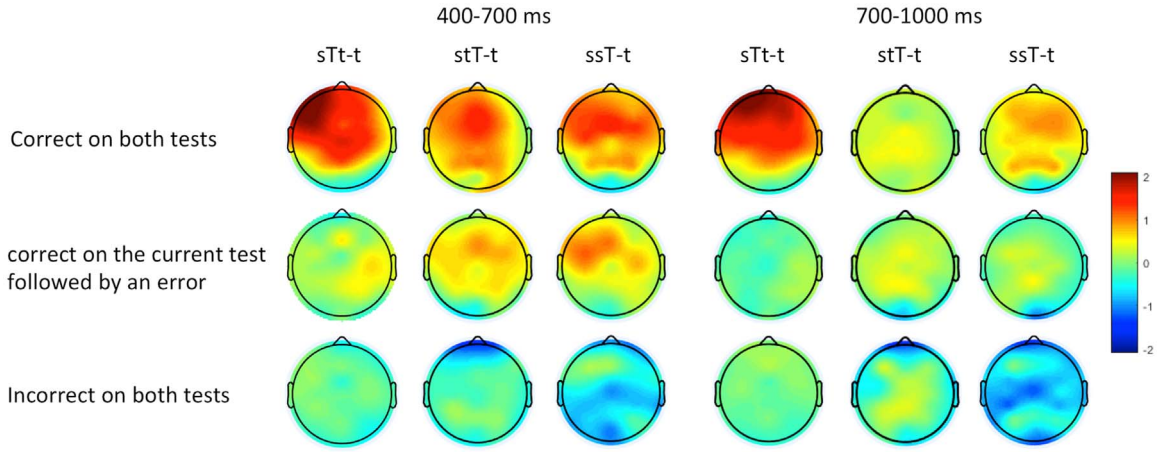

Fig. 4. Head plots illustrating the 400–700 ms and 700–1000 ms time windows for each test phase in each condition as a function of current and subsequent test accuracy. The capital letter indicates the test phase from which the head plots in the corresponding column come from.

test phase as a function of current and subsequent accuracy. We focused on two clusters of electrodes, a frontal cluster (F3, Fz, and F4) and a parietal cluster (P3, Pz, and P4).

Fig. 5 plots the waveforms for the three types of outcome patterns during Phase 2. These are the sTtt condition trials for which the answer was again correct at Phase 3, the trials that switched from correct at Phase 2 to incorrect at Phase 3, and trials that were incorrect at both Phase 2 and Phase 3. Because there was no feedback after retrieval attempts, there were almost no trials that switched from incorrect to correct. We analyzed the ERP components during two time windows. Based on past research (Liu and Reder, 2016; Bridge and Paller, 2012), we propose that the first time window (400–700 ms) reflects the retrieval process and the second time window (700–1000 ms) reflects the re-encoding process.

For the first window (400–700 ms), there was a significant main effect of outcome pattern type, $F(2,50) = 10.82$, $p < .001$, $\eta_p^2 = .30$, but no main effect of electrode clusters nor an interaction between clusters and outcome pattern type, $p > .05$. Amplitudes were more positive for trials for which the answer was correct on tests at both Phase 2 and Phase 3 compared with trials that were correct at Phase 2 but incorrect at Phase 3, $t(25) = 2.22$, $p = .036$, d = .44, and compared with trials that were incorrect at Phase 2 and 3, $t(25) = 4.99$, $p < .001$, d = .98. Likewise, amplitudes were more positive for trials that were correct at Phase 2 but incorrect at Phase 3 than for trials that were incorrect at both Phase 2 and 3, $t(25) = 2.32$, $p = .029$, d = .46. This pattern suggests that the amplitude during the retrieval time window is a good indicator of the quality of the current retrieval and also a good predictor for subsequent test performance. Visual inspection of the waveforms did indeed show that 700 ms was the time point that best distinguished the two effects. At ~700 ms, how well the amplitudes predicted the quality of the current retrieval started to decline. This occurred at ~700 ms at both parietal and frontal locations. However, how well the amplitudes were able to predict subsequent test performance continued to be strong past 700 ms, i.e. during the later time window (700–1000 ms). During the second time window (700–1000 ms), there was a main effect of outcome pattern type on

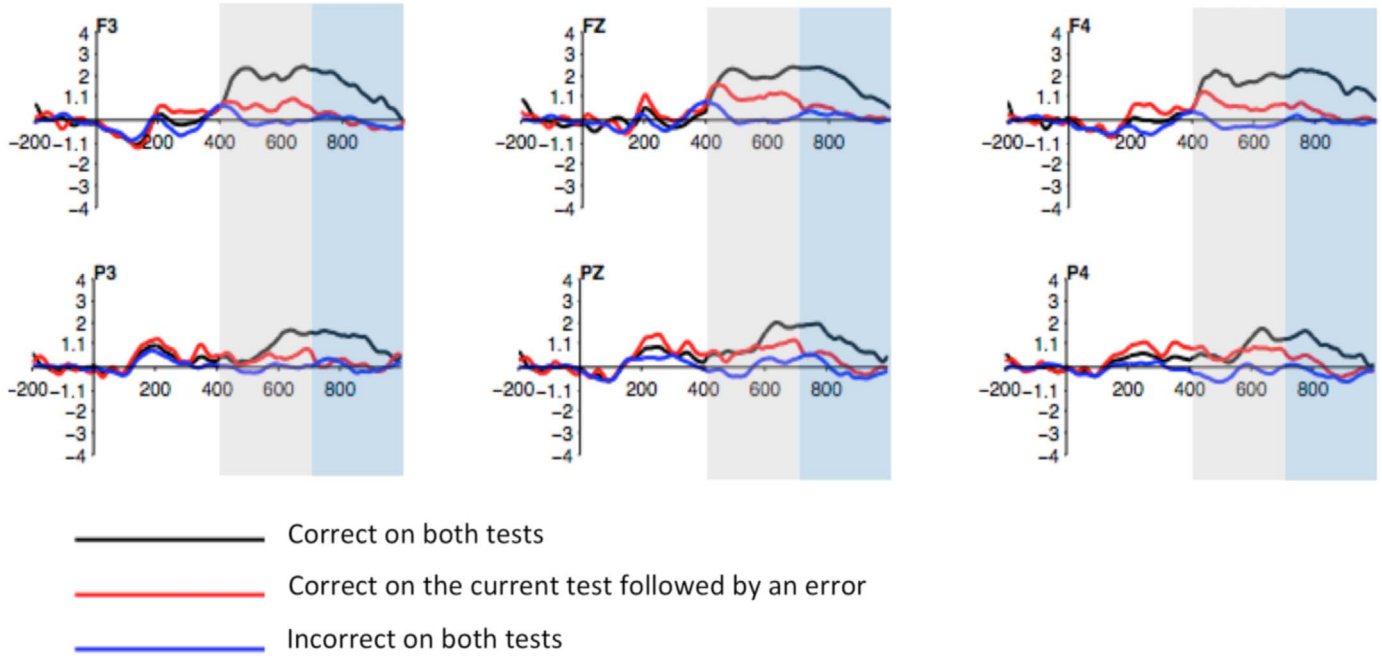

Fig. 5. Waveforms for the first test of the sTtt condition as a function of current and subsequent test accuracy. The two vertical stripes represent the 400–700 ms (gray) and 700–1000 ms (blue) time windows, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

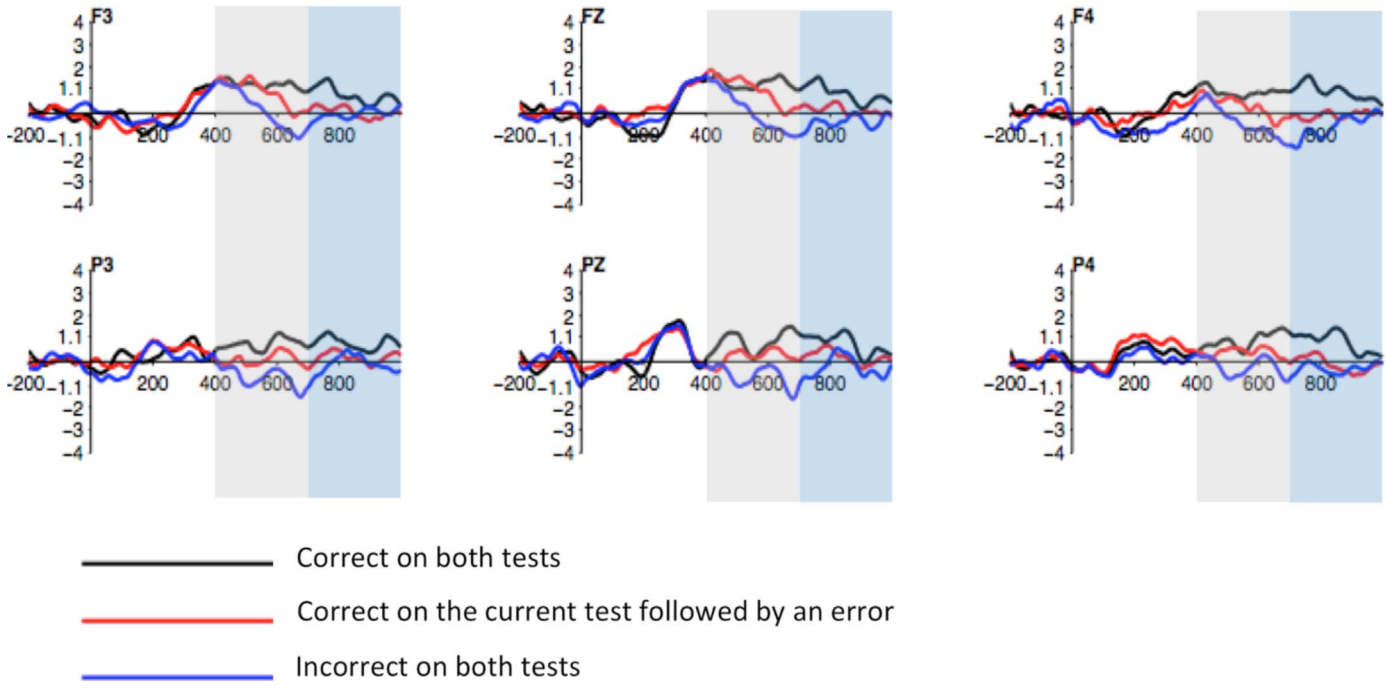

Fig. 6. Waveforms for the first test of the ssTt condition as a function of current and subsequent test accuracy. The two vertical stripes represent the 400–700 ms (gray) and 700–1000 ms (blue) time windows, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

mean amplitude, $F(2,50) = 5.80$, $p = .005$, $\eta_p^2 = .19$. Amplitudes were more positive for trials that were correct at both Phase 2 and Phase 3 than for trials that were correct at Phase 2 but incorrect at Phase 3, $t(25) = 2.1$, $p = .050$, d = .40, and for trials that were incorrect at both Phase 2 and Phase 3, $t(25) = 3.49$, $p = .002$, d = .69. There was no significant interaction between clusters and outcome pattern types, $p > .01$. There was no significant difference between the other two types of trials, $p > .1$. We interpret the amplitude at this time window as an index of the success of re-encoding such that if the re-encoding is strong, the answer will be remembered the next time it is tested.

Fig. 6 shows the waveforms during the Phase 3 test. It examines the ssTt condition and contrasts trials that were correct both for the current (Phase 3) and the subsequent (Phase 4) tests with those trials that were correct on the current test but incorrect on the subsequent test. It also contrasts those that were correct on both tests with trials that were incorrect on both. The patterns of waveforms for the Phase 3 test (ssTt condition) are similar to those observed for the Phase 2 test (sTtt condition). For time window 400–700 ms, there is a main effect of outcome pattern type on mean amplitude, $F(2,50) = 9.31$, $p < .001$, $\eta_p^2 = .27$. Amplitudes were more positive for trials that were correctly answered on both Phase 3 and Phase 4 tests compared with trials that were correct at Phase 3 but wrong at Phase 4, $t(25) = 2.43$, $p = = .022$, d = .48. Amplitudes were also more positive for trials that were correct at Phase 3 but incorrect at Phase 4 than for trials that were incorrect at both Phase 3 and 4, $t(25) = 2.84$, $p = = .009$, d = .56. There was no significant interaction between clusters and outcome pattern types, $p > .01$.

For the time window 700–1000 ms, the effect of outcome pattern type on mean amplitude was also significant, $F(2,50) = 5.14$, $p = .009$, $\eta_p^2 = .17$. Amplitudes were more positive for trials that were correct at both Phase 3 and Phase 4 compared with trials that were correct at Phase 3 but incorrect at Phase 4, $t(25) = 3.61$, $p = .001$, d = .71, and with trials that were incorrect at both Phase 3 and Phase 4, $t(25) = 2.68$, $p = .013$, d = .53. There was no significant difference between the other two types of trials, $p > .1$. The pattern for the two time windows is consistent with the results we observed for the first test in the sTtt condition and support our interpretation that the amplitude at 400–700 ms reflects the quality of the current retrieval and the amplitude at 700–1000 ms is an index of the quality of the post-retrieval re-encoding process. There was no significant interaction between clusters and outcome pattern types, $p > .05$.

3.2.2. *The testing effect involves two processes*

We found that the amplitudes at the time window 400–700 ms differed as a function of current accuracy, with more positive amplitudes associated with better performance on the current test. We interpret the amplitude of this component as reflecting the strength or amount of the memory trace retrieved. This result is consistent with previous ERP studies examining ERP components associated with successful memory retrieval (e.g., Bridge and Paller, 2012; Allan and Rugg, 1997; see Wilding and Ranganath, 2012 for a review). Two components that are commonly associated with successful retrieval are the FN400 and the parietal old-new effect (Rugg and Curran, 2007; see Yonelinas, 2002 for a review). Several studies have found evidence suggesting that the FN400 is an index of familiarity (e.g. Curran, 2000, Rugg and Curran, 2007). However, our study was concerned with cued-recall and, as such, subjects needed to rely on recollection, not familiarity. Therefore, we did not expect to see an effect earlier than 400 ms. Consistent with this prediction, there were no significant effects of retrieval outcomes during the 200–400 ms time window. On the contrary, the parietal old-new effect is often explained as an index of the amount of information recollected (Vilberg et al., 2006). While our explanation of the early time window (400–700 ms) is consistent with the literature on the parietal old-new effect, we did not find significant interactions between retrieval outcomes and locations. Furthermore, visual inspection of the scalp maps suggested that the effect was distributed over both frontal and parietal regions. This is consistent with prior ERP studies that compared memory retrieval in recognition and cued-recall tasks (e.g., Allan and Rugg, 1997). They found that although the onset latencies of the parietal old-new effect in a recognition task and retrieval success effect in a cued-recall task are similar, the cued-recall effect is more diffusely distributed over the scalp than the recognition old/new effect.

Additionally, we found that the amplitude of this component also predicted subsequent memory performance, with more positive amplitudes associated with better subsequent performance. This seems

logical if a more positive amplitude reflects a stronger memory trace retrieved. The subsequent memory effect was distributed over both frontal and parietal regions. This pattern is also consistent with our prior fMRI studies (Liu et al., 2014; Liu and Reder, 2016) showing that brain activity in both frontal and parietal regions during memory retrieval could predict subsequent memory performance.

The amplitudes during the second time window at 700–1000 ms, were shown to predict success on a subsequent test but those amplitude differences were not related to current accuracy. We had postulated that this window reflects a memory re-encoding process and the pattern of waveforms support that hypothesis. This pattern is also consistent with Bridge and Paller (2012) in which they found that differences in amplitude during the second time window reflect differences in the effort of re-encoding what was retrieved (even if incorrectly). These results are also consistent with another ERP study using a similar paradigm (Bai et al., 2015). Although their primary analyses focused on the differences between test and re-study conditions, they also found that a late component (700–1000 ms) only predicted subsequent test performance, without being correlated with current accuracy. In other words, these results support the view that both a retrieval process and a memory re-encoding process underlie the benefits afforded by testing, and that they occur sequentially.

3.2.3. Subsequent memory effects based on ERPs during the second test

If these two time windows do indeed reflect two distinct processes, then we might want to see more dissociations, in addition to the one described in the previous section (i.e., that the first time window predicts both current and subsequent accuracy, while the second only predicts subsequent accuracy). Our previous research (Liu and Reder, 2016) suggested that these two processes are differentially affected by overlearning. Thus, we now examine whether the waveforms associated with the first and second time window show similar or different patterns based on the degree of overlearning.

In Liu and Reder's (2016) fMRI study, the two processes were distinguished by examining the different patterns of brain activity rather than by examining time windows. We found that the brain regions uniquely associated with retrieval, rather than encoding, right PFC and right PPC, were found to always predict subsequent memory success while, the regions also associated with encoding, left PFC and left PPC, were not always predictive. Specifically, when the information being tested had already been successfully recalled several times, activation in the left PFC and PPC were no longer predictive of subsequent recall.

We believe that when a retrieval attempt is successful, the retrieval process necessarily contributes to subsequent successful retrievals by virtue of strengthening and building the retrieval paths. On the other hand, the post-retrieval, re-encoding process can disengage when the recalled information is already well learned (e.g., when it has already been correctly recalled twice). In other words, the re-encoding process is influenced by a sense of novelty such that a well-learned answer would not warrant attention from the process.

In order to investigate whether the extent of the re-encoding process is diminished when the information has already been well learned, we examined the second test in the stTt condition for trials that had been correctly recalled twice (shown in Fig. 7).

The mean amplitudes at 400–700 ms for the stTt condition (second test) showed a very similar pattern to what was observed for the first test in the sTtt condition. Specifically, again there was a significant main effect of outcome pattern type, $F(2,48) = 7.53$, $p = .001$, $\eta_p^2 = .24$, and amplitudes were more positive for trials for which the answer was correct at both Phase 3 and 4 compared to trials that were correct at Phase 3 but incorrect at Phase 4, $t(24) = 3.24$, $p = .003$, d = .65. Amplitudes were also more positive for trials that were correct at Phase 3 but incorrect at Phase 4 than for trials that were incorrect at both Phase 3 and 4, $t(24) = 3.47$, $p = .002$, d = .69. There was no significant interaction between clusters and outcome pattern types, $p > .05$.

On the other hand, and most importantly, the waveform pattern at the second time window, 700–1000 ms, looked quite different from that observed during the first test phase in both the sTtt and the ssTt conditions. During the first test in the sTtt (Fig. 5) and the ssTt (Fig. 6) conditions, the amplitudes during the second time window predicted the subsequent memory performance. However, during the second test in the stTt condition, when information has been correctly retrieved twice, there was no significant effect of outcome pattern type or interaction between clusters and outcome pattern types, $p > .1$. This replicates our finding from an fMRI version of this experiment (Liu and Reder, 2016). There and here, we interpret these results as suggesting that while the retrieval process consistently contributes to the learning process through testing, the re-encoding process becomes disengaged when the answers have been well learned. This explanation is also in line with the literature on the post-retrieval monitoring effect, which is shown to be an index of post-retrieval assessment of retrieved information (see Wilding and Ranganath, 2012 for a review). Prior studies suggested that this effect only emerges when post-retrieval assessment of retrieved information is required (Wilding and Rugg, 1996; Hayama et al., 2008).

4. General discussion and conclusion

The current study used ERPs to test our hypothesis that there are two sequential processes that underlie the benefits of testing, namely memory retrieval and re-encoding, in that order. Our results indicate that the ERP component associated with retrieval (400–700 ms) predicts current and subsequent accuracy. In contrast, the component that follows it, 700–1000 ms, only predicts subsequent accuracy. Moreover, while the retrieval time window (400–700 ms) predicts subsequent test accuracy for both the second and third test, the re-encoding time window (700–1000 ms), only predicts subsequent memory from the first to the second test but not from the second to the third test. Our interpretation is that when the information has been well learned as in the case of two successive correct recalls, there is little need for re-encoding. This dissociation between the two processes was also found in a similar paradigm using fMRI (Liu and Reder, 2016).

If the testing effect is superior to re-study, why did we not find an advantage of testing in this experiment? We have argued here that testing is better than re-study because subjects get practice at retrieval as well as an additional opportunity to re-encode the retrieved information. On the other hand, for items that are not learned well enough to be retrieved on the initial test, since there is no feedback, there is no chance to practice retrieval or re-encode the information. In contrast, in the re-study condition subjects are given the opportunity to re-encode all the items. It has been previously established that the testing effect is modulated by whether feedback is provided after retrieval and the difficulty of initial acquisition (e.g., Rowland, 2014). When feedback is given after every test trial or when feedback is given until at least there is one correct retrieval, the testing effect is almost always superior after a moderate delay from the intervening test to the final test. If feedback is not provided then the learning outcome from a re-study practice will be inferior to that from items that are correctly answered on test practice, but the average results will depend on the overall difficulty of the items to be acquired. We chose not to provide feedback in this study so that we could more carefully examine the contribution of the re-encoding process from retrieval without contamination from feedback.

In summary, our earlier research motivated the novel hypothesis that there are two separate processes, which occur sequentially, that underlie the memorial advantages of the testing effect, and the present study provided essential converging evidence to support this view. By using ERP methodology, we were able to explicitly examine the two separate processes over time: First, a retrieval process that strengthens the associated links that bring the answer to mind and, second, a re-encoding process that affords an additional encoding opportunity of the

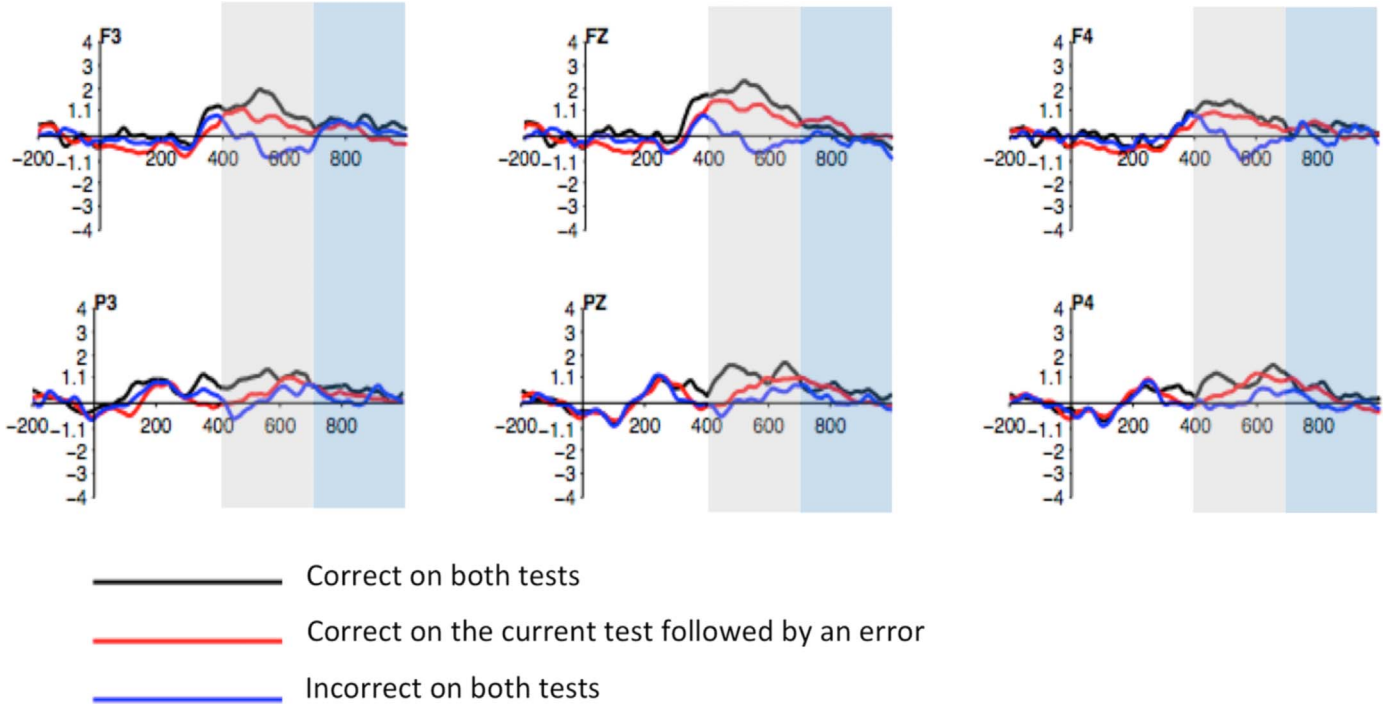

Fig. 7. Waveforms for the second test of the stTt condition as a function of current and subsequent test accuracy. The two vertical stripes represent the 400–700 ms (gray) and 700–1000 ms (blue) time windows, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

correctly retrieved answer. By employing converging measures such as ERP, our results help shed light on the mechanisms involved in learning from tests in a way that would not have been possible from behavioral and fMRI studies alone.

Acknowledgment

Funding: This work was supported by National Natural Science Foundation of China (31700996), Chinese Ministry of Education, Humanities and Social Sciences, Youth Fund (16YJC190015) and the Fundamental Research Funds for the Central Universities 1350-ZK1013).

We thank Ya Zhang, Michelle Stepan and Shanshan Dong for assistance in running subjects and data analyses and John Anderson and Lori Holt for commenting on previous drafts of the manuscript.

References

Allan, K., Rugg, M.D., 1997. An event-related potential study of explicit memory on tests of cued recall and recognition. Neuropsychologia 35 (4), 387–397.
Anderson, J.R., Reder, L.M., 1979. An elaborative processing explanation of depth of processing. In: Cermak, L.S., Craik, F.I.M. (Eds.), Levels of Processing in Human Memory. Erlbaum, Hillsdale, NJ.
Bai, C., Bridger, E.K., Zimmer, H.D., Mecklinger, A., 2015. The beneficial effect of testing: an event-related potential study. Front. Behav. Neurosci (248-248).
Bridge, D.J., Paller, K.A., 2012. Neural correlates of reactivation and retrieval-induced distortion. J. Neurosci. 32 (35), 12144–12151.
Carpenter, S.K., 2009. Cue strength as a moderator of the testing effect: the benefits of elaborative retrieval. J. Exp. Psychol. Learn. Mem. Cogn. 35 (35), 1563–1569.
Carpenter, S.K., Delosh, E.L., 2006. Impoverished cue support enhances subsequent retention: support for the elaborative retrieval explanation of the testing effect. Mem. Cogn. 34 (2), 268–276.
Coltheart, M., 1981. The MRC psycholinguistic database. Q. J. Exp. Psychol. 33 (4), 497–505.
Curran, T., 2000. Brain potentials of recollection and familiarity. Mem. Cogn. 28 (6), 923–938.
Delorme, A., Makeig, S., 2004. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J. Neurosci. Methods 134 (1), 9–21.
Den Broek, G.S., Takashima, A., Segers, E., Fernandez, G., Verhoeven, L., 2013. Neural correlates of testing effects in vocabulary learning. NeuroImage 94–102.
Dudai, Y., 2004. The neurobiology of consolidations, or, how stable is the Engram? Annu. Rev. Psychol. 51–86.
Finn, B., Roediger, H.L., 2011. Enhancing retention through reconsolidation negative emotional arousal following retrieval enhances later recall. Psychol. Sci. 22 (6), 781–786.
Griffin, M., Dewolf, M., Keinath, A., Liu, X.L., Reder, L.M., 2013. Identical versus conceptual repetition FN400 and parietal old/new ERP components occur during encoding and predict subsequent memory. Brain Res. 68–77.
Hayama, H.R., Johnson, J.D., Rugg, M.D., 2008. The relationship between the right frontal old/new ERP effect and post-retrieval monitoring: specific or non-specific? Neuropsychologia 46 (5), 1211–1223.
Hogan, R.M., Kintsch, W., 1971. Differential effects of study and test trials on long-term recognition and recall. J. Verbal Learn. Verbal Behav. 10 (5), 562–567.
Johansson, M., Mecklinger, A., 2003. The late posterior negativity in ERP studies of episodic memory: action monitoring and retrieval of attribute conjunctions. Biol. Psychol. 91–117.
Kang, S.H., Mcdermott, K.B., Roediger, H.L., 2007. Test format and corrective feedback modify the effect of testing on long-term retention. Eur. J. Cogn. Psychol. 528–558.
Karpicke, J.D., Lehman, M., Aue, W.R., 2014. Chapter seven – retrieval-based learning: an episodic context account. Psychol. Learn. Motiv. 237–284.
Karpicke, J.D., Roediger, H.L., 2008. The critical importance of retrieval for learning. Science 319 (5865), 966–968.
Kim, H., 2011. Neural activity that predicts subsequent memory and forgetting: a meta-analysis of 74 fMRI studies. NeuroImage 54 (3), 2446–2461.
Kornell, N., Bjork, R.A., Garcia, M.A., 2011. Why tests appear to prevent forgetting: a distribution-based bifurcation model. J. Mem. Lang. 65 (2), 85–97.
Liu, X.L., Reder, L.M., 2016. Fmri exploration of pedagogical benefits of repeated testing: when more is not always better. Brain Behav. 6, 7.
Liu, X.L., Liang, P., Li, K., Reder, L.M., 2014. Uncovering the neural mechanisms underlying learning from tests. PLoS One 9 (3), e92025.
Lopez-Calderon, J., Luck, S.J., 2014. Erplab: an open-source toolbox for the analysis of event-related potentials. frontiers in human. Neuroscience 8 (3), 213.
Pyc, M.A., Rawson, K.A., 2009. Testing the retrieval effort hypothesis: does greater difficulty correctly recalling information lead to higher levels of memory? J. Mem. Lang. 60 (4), 437–447.
Roediger, H.L., Butler, A.C., 2011. The critical role of retrieval practice in long-term retention. Trends Cogn. Sci. 15 (1), 20–27.
Roediger, H.L., Karpicke, J.D., 2006a. Test-enhanced learning taking memory tests improves long-term retention. Psychol. Sci. 17 (3).
Roediger, H.L., Karpicke, J.D., 2006b. The power of testing memory basic research and implications for educational practice. Perspect. Psychol. Sci. 1 (3), 181–210.
Rosburg, T., Johansson, M., Weigl, M., Mecklinger, A., 2015. How does testing affect retrieval-related processes? - An event-related potential (ERP) study on the short-term effects of repeated retrieval. Cogn. Affect. Behav. Neurosci. 15 (1), 195–210.
Rowland, C.A., 2014. The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. Psychol. Bull. 140 (6), 1432–1463.
Rowland, C.A., Delosh, E.L., 2015. Mnemonic benefits of retrieval practice at short retention intervals. Memory 23 (3), 403–419.
Rugg, M.D., Curran, T., 2007. Event-related potentials and recognition memory. Trends

Cogn. Sci. 11 (6), 251–257.
Spitzer, B., Hanslmayr, S., Opitz, B., Mecklinger, A., 2009. Oscillatory correlates of retrieval-induced forgetting in recognition memory. J. Cogn. Neurosci. 21 (5), 976–990.
Toppino, T.C., Cohen, M.S., 2009. The testing effect and the retention interval: questions and answers. Exp. Psychol. 56 (4).
Vilberg, K.L., Moosavi, R.F., Rugg, M.D., 2006. The relationship between electrophysiological correlates of recollection and amount of information retrieved. Brain Res. 1122 (1), 161–170.
Wagner, A.D., Schacter, D.L., Rotte, M., Koutstaal, W., Maril, A., Dale, A.M., Rosen, B.R., Buckner, R.L., 1998. Building memories: remembering and forgetting of verbal experiences as predicted by brain activity. Science 281 (5380), 1188–1191.
Wheeler, M.A., Ewers, M., Buonanno, J., 2003. Different rates of forgetting following study versus test trials. Memory 11 (6), 571–580.
Wheeler, M.S., Roediger, H.L., 1992. Disparate effects of repeated testing: reconciling Ballard's (1913) and Bartlett's (1932) results. Psychol. Sci. 3 (4), 240–245.
Wilding, E.L., Ranganath, C., 2012. Electrophysiological correlates of episodic memory processes. In: Luck, S.J., Kappenman, E.S. (Eds.), The Oxford Handbook of Event-related Potential Components. Oxford University Press, New York.
Wilding, E.L., Rugg, M.D., 1996. An event-related potential study of recognition memory with and without retrieval of source. Brain 119 (3), 889–905.
Wing, E.A., Marsh, E.J., Cabeza, R., 2013. Neural correlates of retrieval-based memory enhancement: an fMRI study of the testing effect. Neuropsychologia 51 (12), 2360–2370.
Yonelinas, A.P., 2002. The nature of recollection and familiarity: a review of 30 years of research. J. Mem. Lang. 46 (3), 441–517.